International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

Credit Card Fraud Detection Using Bagging and Boosting Algorithms

Akansha Thakarke[1], Sakshi Ugale[2], Sneha Nale[3], Dr. Mrudul Dixit[4]

[1]Electronics and Telecommunication, MKSSS's Cummins College of Engineering for Women, Pune, India
[1]*akanshathakarke[at]gmail.com*
[2]*sakshiugale200024[at]gmail.com*
[3]*nale.sneha10[at]gmail.com*
[4]*mrudul.dixit[at]cumminscollege.in*

Abstract: ***The term Credit Card Fraud indicates that the defrauder is using your credit card credentials or has stolen your credit card for his/her financial benefit. Research tells us that due to economic expansion in recent years, credit card spending has increased. This eventually leads to increase in fraudulent credit card transactions. In the last few years, this has been a predominant issue; it causes a huge loss to the companies and the cardholder. The paper talks about machine learning techniques such as Logistic Regression, Naïve Bayes, Boosting Classifier and Bagging classifier to detect credit card frauds.***

Keywords: Bagging, Boosting, Credit Card fraud detection, Decision tree, Ensemble learning

1. Introduction

Recently, there has been a major shift when it comes to payments. People are opting for digital payment methods such as credit/debit cards, net banking, and Unified Payment Interface (UPI) methods. The usage of credit cards is expected to grow by 25% as predicted by the experts. Hence, it is crucial for the banks to provide a secure payment interface. Machine learning algorithms have been boon to the banks for safe and reliable techniques to detect fraud transactions. Credit card frauds are unauthorised transactions when the owner is unaware of the fact that the card is being used incorrectly. The paper consists of computations of various machine learning algorithms to perceive the most accurate technique to detect frauds. Primarily, these algorithms can be implemented in Python and R language. However, Python has dominance overR because of the numerous existing libraries. The dataset of credit card transactions is fetched from Kaggle.

2. Literature Survey

[1]Y. Sahin and E. Duman performed a research in 2011, where they have used seven Machine Learning classification model for detection in credit card fraud. ANN has enhanced performance as compared to Logistic regression.[2] Random Forest algorithm is used for detection of fraudulent transactions of credit card. The result obtained in this shows that Random forest gives about 90% of accuracy. [3]Kuldeep Randhawa 1, Chu Kiong Loo, has applied standard models like SVM, Naive Bayes as well as Hybrid models like Adaboost and Majority Voting combination methods. The performance is determined on the basis of MCC score. Majority Voting has achieved the best MCC score. [4] Admel Husejinović determined best performing algorithm amongst Naive bayes, C4.5 decision tree and bagging DECISION machine learning algorithms on the basis of more precise results by calculating PRC area, and PRC rates. [5]Rashmi S. More, Chetan J. Awati, Dr. Suresh K. Shirgave, Dr. Rashmi J. Deshmukh, Sonam S. Patil reviewed various ML algorithms and have selected Random Forest Classifier which is trained using feedback and delayed supervised sample, further they have processed learning to rank approach where frauds will be ranked based on priority to solve class imbalance and concept drift problem. [6]S P Maniraj, Aditya Saini, Swarna Deep Sarkar, Shadab Ahmedhave listed most common methods of fraud and its detection methods. Moreover, the paper discusses how to apply machine learning algorithms to get better results and also comments on high percentage of accuracy and low precision. [7] Rahul Goyal, Amit Kumar Manjhvar, Vikas Sejwar implemented Logistic regression and XGBoost to conduct static learning method and incremental learning method. Obtained precision and accuracy higher in case of XGBoost algorithm also ROC was increased to 0.3. [8]Lakshmi S V S S, Selvani Deepthi Kavila used Logistic regression, decision tree and random forest to detect fraudulent transactions, calculated accuracy, precision sensitivity and specificity to evaluateperformance of the system. By comparing all the results Random Forest is better than other two.

3. Problem Definition

Credit Card Fraud Detection problem means to create a ML model and train it with the dataset of past credit card transactions which comprises the knowledge about which cases are fraud and which are genuine and then using this model to find out if the new transaction is fraud or not.

4. Methodology

Figure 1:System Architecture

Figure 1 shows the steps to solve the credit card fraud detection problem.The dataset is divided in two parts – training (80%) and testing (20%) data.This dataset is then balanced using SMOTE function. The machine learning model is then trained with training dataset. The trained model is then used to differentiate between Fraud and Genuine transactions from the test data.

4.1 Dataset Collection

The dataset used is taken from kaggle. The total number of transactions present in the dataset are 2, 84,807. Except for 'Time' and 'Amount' all the features are transformed with PCA for confidentiality reasons.

4.2 Data Preprocessing

Data pre-processing is an integral step in machine learning as the quality of data and the information achieved from it directly affects the ability of machine learning model to learn; therefore, it is quiteimportant topre-process the data before training a machine learning model.

Pre-processing steps include:

Handling Missing Data: Missing data can cause errors in the Machine Learning model. No Missing data is present in this dataset

Encoding Categorical Data: Every string is encoded to 1, 2, 3 …… or one column is turned into separate columns.

Dataset Splitting: The dataset is split using the 'train_test_split' module into80% training and 20% testing dataset.

Feature scaling: It puts all the features on the same scale. It can be done using Standardisation or Normalisation methods.

Data Imbalance: Figure 2 shows highly imbalanced as only 492(0.17%) fraud cases are present of the total dataset. This can be an issue in finding the accuracy of the model. Hence it is important to Balance the data properly.

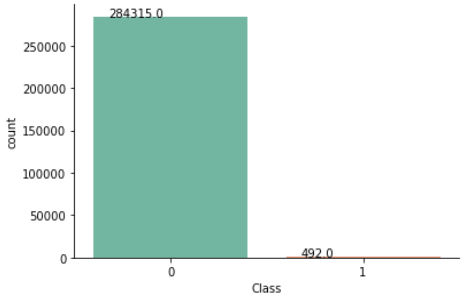

Figure 2: Imbalance Dataset

The dataset can be balanced using different techniques like under sampling Majority Class, Oversampling Minority Class. Here SMOTE function is used to balance the data. Figure 3 shows the fraud and genuine cases after balancing.

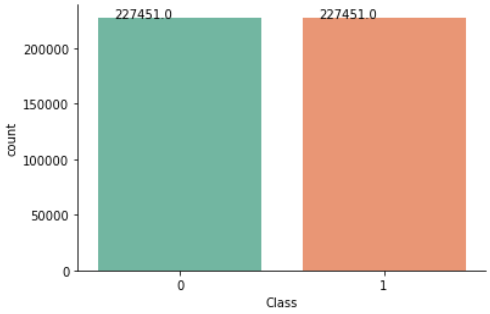

Figure 3: Balanced Dataset

4.3 Feature Selection

There are total 30 features (columns) present in the dataset. Not all of these features contribute much to the classifier model. Through Feature Selection, we can remove these redundant features which will lead to boosting of the performance of the classifier model. Different techniques are available for feature selection like Feature importance, Correlation Matrix using heat map, Univarate Selection. Density distribution is used to drop the 'time' feature as it does not contribute much to the output.

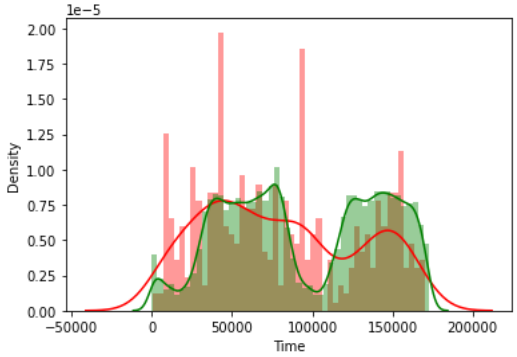

Figure 4: Density distribution of Time Feature

As this feature doesn't contribute much to differentiate between Fraud and Genuine Transactions, it can be dropped. It helps to improve the predictions of the classifier.

5. Machine Learning Algorithms Implemented

In supervised Learning 'labeled' data is used for training the model. Input data is provided for the training along with the output.In Unsupervised Learning data used for training the model is not labelled. No output data is provided to the model in advance.

5.1 Standard Models

a) Logistic Regression
Logistic Regression uses Cost Function to anticipate the probability of an outcome on the basis of numerous predictors which include numerical and categorical functions.

b) Naive Bayesian
Naive Bayes Algorithm is based on Bayes' Theorem. It is a supervised learning classification technique which assumes that the occurrence of each feature is independent of another feature. This makes the algorithm very fast as compared to other complicated algorithms.

5.2 Ensemble Learning

In Ensemble Learning, multiple Weak Learners are combined to form a more complex model which will have better accuracy than the individual model.

a) Bagging
Bootstrapping Aggregation also known as Bagging is used for complex problems in Machine learning to enhance the precision and accuracy of the model.

b) Boosting
In the Boosting Model, weak and imprecise (inaccurate) rules are integrated to achieve more accurate and precise Prediction. This results in reducing Variance and Bias. in supervised learning.

Adaboost uses serial weak learners to solve non-linear problems; this is achieved by increasing the weight of fallacious decisions and vice versa. This can be a single layer perceptron.

Extreme Gradient Boosting Machine (XGBoost) is a modified model of the GBM algorithm. XGBoost builds its trees sequentially to rectify ill-prediction of the previous trees. To make this process faster, it performs parallel preprocessing. It can control the missing values in the dataset autonomously.

A. Stacking
In Stacking Ensemble, outcomes of various Machine Learning algorithms are used as input in second layer learning algorithms. This algorithm is trained to combine the model predictions and form a new set of predictions.

6. Result and Analysis

Confusion Matrix describes the performance of a classification model. Figure 4 is the confusion matrix. The results of the machine learning algorithms are evaluated on the basis of accuracy precision and recall values derived from the confusion matrix.

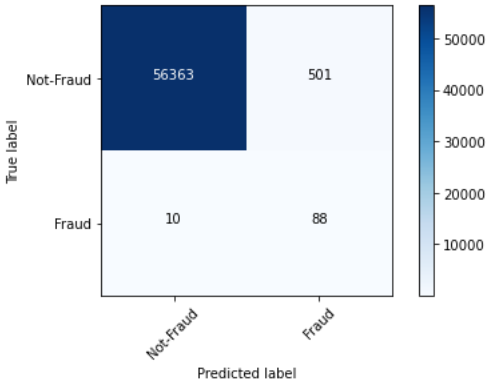

Figure 5: Confusion matrix

The following tablesshows performance of all the algorithms.

Table 1: Standard Algorithms Performance in %

Algorithm	Accuracy	Precision	Recall
Naïve Bayes	97.29	5.31	87.75
Logistic Regression	97.79	7.22	90.81
Decision Tree	99.79	44.82	79.59

Table 2: Boosting Classifier Performance in %

Algorithm	Accuracy	Precision	Recall
Adaboost	98.86	12.24	90.81
LGBM	99.48	23.57	88.77
XGBM	99.17	16.27	91.83

Table 3: Bagging Classifier Performance in %

Algorithm	Accuracy	Precision	Recall
Random Forest	99.94	86.17	82.65
Bagging	99.89	66.39	82.65

7. Conclusion

Standard machine learning models like Naïve Bayes, Logistic Regression, Decision Tree and Hybrid machine learning models such as Boosting and Bagging have been implemented for Credit Card fraud detection. Considering parameters like Accuracy, Precision, Recall; among the standard models, Decision Tree classifier has been determined to be more suited for fraud detection. Similarly in Boosting Classifier, LGBM Model over performs Adaboost and XGBM. Bagging ensemble model is found to have best results as compared to standard and boosting algorithms. Therefore, Bagging Ensemble Techniques can be considered as a ML technique for Credit Card Fraud Detection.

8. Future Scope

The highest accuracy is attained by the Decision Tree model for credit card fraud detection of 99.79%. This is the maximum value that the machine learning algorithms could pull off as we have tried to execute several machine learning algorithms. The accuracy can be enhanced by combining various algorithms together. Nonetheless, the output should be in correspondence to that of the other outputs obtained. Hence, this project enables us to explore and incorporate algorithms to yield desirable results. The model can be strengthened by normalizing the dataset with more efficient standardisation techniques. In order to maximize precision, a dataset with larger size can be implemented as well. The more data available to train and test, the more will be the accuracy of predicting frauds in less time.

References

[1] Sahin, Yusuf & Duman, Ekrem. (2011). Detecting Credit Card Fraud by Decision Trees and Support Vector Machines. IMECS 2011 - International MultiConference of Engineers and Computer Scientists 2011. 1. 442-447.
[2] M.Suresh Kumar, V.Soundarya, S.Kavitha . E.S.Keerthika ,E.Aswini “Credit card fraud detection using Random Forest algorithm” ICCCT 2019.
[3] Kuldeep Randhawa, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim, Asoke K. Nand, “Credit Card Fraud Detection using AdaBoost and Majority Voting, 2018
[4] Admel Husejinovic, “Credit card fraud detection using naive Bayesian and C4.5 decision tree classifiers”, 2020
[5] Rashmi S. More, Chetan J. Awati, Dr. Suresh K. Shirgave, Dr. Rashmi J. Deshmukh, Sonam S. Patil, “Credit Card Fraud Detection using Supervised Learning Approach”, International Journal Of Scientific & Technology Research Volume 9, Issue 10, October 2020.
[6] S P Maniraj, Aditya Saini, Swarna Deep Sarkar, Shadab Ahmed, “Credit Card Fraud Detection using Machine Learning and Data Science”, International Journal of Engineering Research and Technology (IJERT), ISSN: 2278-0181, Vol. 8 Issue 09, September-2019
[7] Rahul Goyal, Amit Kumar Manjhvar, Vikas Sejwar, “Credit Card Fraud Detection in Data Mining using XGBoost Classifier”, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-9 Issue-1, May 2020
[8] Lakshmi S V S, Selvani Deepthi Kavila, “Machine Learning For Credit Card Fraud Detection System”, International Journal of Applied Engineering Research

Author Profile

Akansha Thakarke, third year, Electronics and Telecommunication Engineering from Cummins College of Engineering for Women, Pune. Have a keen interest in Artificial Intelligence and Machine Learning and how these can be used to solve societal issues and real world industry problems.

Sakshi Ugale, third year, Electronics and Telecommunication Engineering from Cummins College of Engineering for Women, Pune. Interested in Data Analytics and Data Science.

Sneha Nale, final year, Electronics and Telecommunication Engineering from Cummins College of Engineering for Women, Pune. Interested in Data Analytics and Machine learning.

Dr. Mrudul Dixit, Dean Alumni and Assistant Professor Cummins College of Engineering for Women, Pune. 21 Years of teaching experience. Published over 35 papers in national, international conferences and journals. Area of specialization: AI, ML, Computer networks and security.